

# Bayesian Hierarchical MPT Modeling Theory

Daniel W. Heck



2023-04-23

# Bayesian Hierarchical MPT Models

- 1] MPT models & heterogeneity
- 2] Hierarchical MPT models
- 3] Bayesian estimation with MCMC sampling
- 4] Advantages of MCMC
- 5] Application: Linking covariates to MPT parameters

## MPT models & heterogeneity

## Standard MPT models assume that ...

- ... people behave identically
- ... items are similarly difficult
- Technical assumption
  - Fixed-effects model: Observations are “independent and identically” (i.i.d.) distributed
  - The likelihood of all observations  $i = 1, \dots, n$  is the product of the likelihood of a single observation  $x_i$

$$p(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^n p(x_i \mid \theta)$$

What about real data?

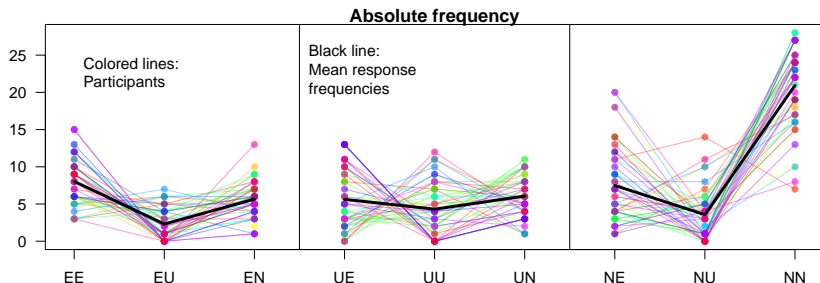
# People Behave Differently

## Source-monitoring task

- 1) Study phase: List of words from Source A and B.
- 2) Test phase: Is the presented item from Source A/B/New?

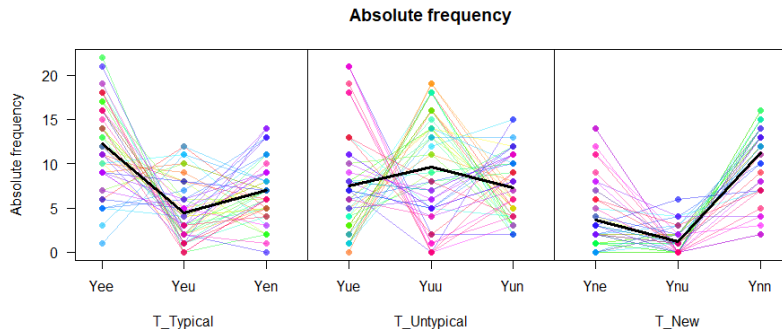
## Distribution of individual response frequencies

- Example: Experiment on schema activation (Arnold et al., 2013)



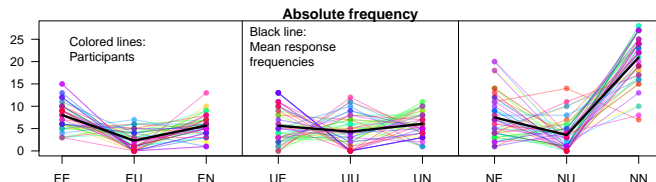
# People Behave Differently

## Data from a different experiment (Bayen, 2011)



- Substantial variance in the choice patterns of participants
  - Differences in memory? Response bias?
- If we fit a standard MPT model to the aggregated data, these differences are ignored (treated as random, unsystematic noise)

# People Behave Differently



## Heterogeneity of participants

- Response frequencies are often aggregated across subjects
  - Dependent variable: Summed individual frequencies
- However, responses are likely not i.i.d.
  - Assumption can be tested statistically (Smith & Batchelder, 2008)
- Heterogeneity may result in biased statistical inference
  - Biased point estimates if parameter are correlated
  - Over-/underestimation of confidence intervals
  - Inflated model-fit statistics

# How to Handle Heterogeneity?

- A) **Complete pooling:** Analysis of aggregated frequencies
  - Ignores differences between persons
  - High power, but possibly biased statistical inference
- B) **No pooling:** A separate MPT model per person
  - Low power, parameter estimates will have a large variance
  - Often, not enough data per participant
  - Problem: How to aggregate results across models?
- C) **Partial pooling:** Hierarchical model
  - Account for differences AND similarities between persons jointly
  - Higher efficiency than separate analysis
  - Individual and group-level parameters inform each other

Note: This classification is very general and not limited to MPT models.



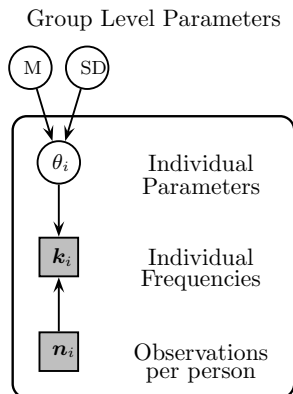
## Hierarchical MPT models

# Hierarchical MPT Models

## Bayesian hierarchical MPT

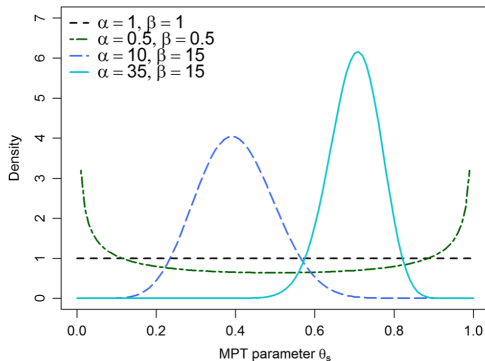
(Klauer, 2010; Smith & Batchelder, 2010)

- Explicit model for participant heterogeneity
- Assumption: MPT structure holds for each person, but with different parameters!
- One parameter vector  $\theta_i = (D_i, d_i, g_i, \dots)$  per person
- On the group level, the  $\theta_i$  have a specific distribution
  - A Beta-MPT: Beta distribution
  - B Latent-trait MPT: multivariate normal distribution for the probit-transformed parameters



## Beta distribution

- Ideally suited to model the distribution of an MPT parameter:
  - Allows values between 0 and 1
  - Two shape parameters:  $\alpha$  and  $\beta$
- On the group level, the mean for the MPT parameter equals:  $\alpha/(\alpha + \beta)$



## Beta-MPT (Smith & Batchelder, 2010)

Parameters:

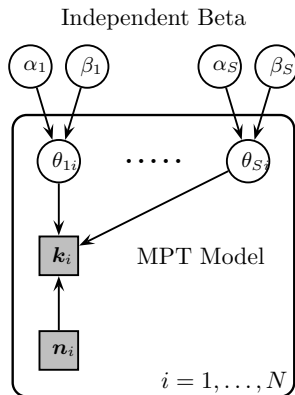
- Level-1: MPT parameters  $\theta_{si}$  of person  $i$
- Level-2: Shape parameters  $\alpha_s$  and  $\beta_s$  of beta distributions

Data:

- $k_i$ : Individual choice frequencies
- $n_i$ : Number of responses per person

Priors:

- Uniform or gamma on  $\alpha_s$  and  $\beta_s$
- Truncation to  $\alpha_s \geq 1$  and  $\beta_s \geq 1$ :  
Unimodal group-level distribution



# Latent-Trait MPT

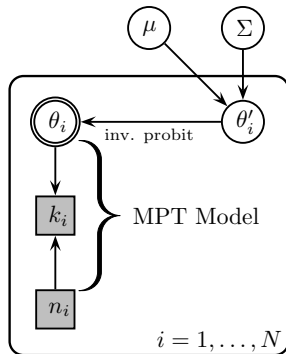
## Latent-trait MPT (Klauer, 2010)

Parameters:

- Level-1: Person parameters are probit-transformed
  - $\theta_{si} = \Phi(\theta'_{si})$
  - $\Phi$  = cumulative density function of the standard normal
- Level-2: Probit-transformed parameters have a multivariate normal distribution
  - Mean  $\mu$  and covariance matrix  $\Sigma$  (on probit scale)

Prior distributions:

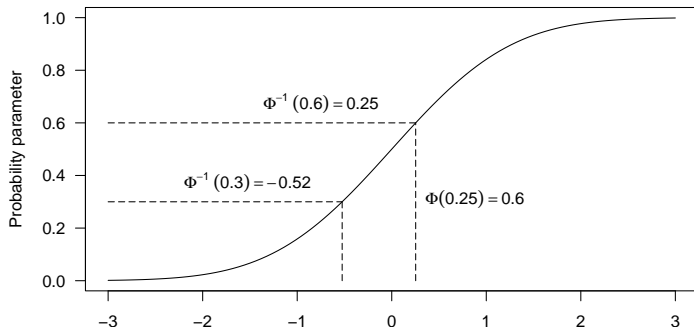
- Standard normal distributions for  $\mu$
- Scaled inverse-Wishart prior for  $\Sigma$



# The Probit-Transformation

## Transformation of MPT parameters

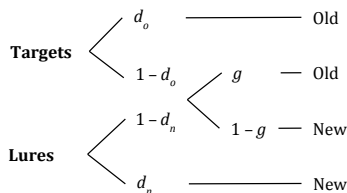
- We need to transform the probability parameters ( $d, g, \dots$ )
- We want parameters between  $(-\infty, +\infty)$  (to work with normal distributions)
- Solution: Transform parameters using the cumulative density function  $\Phi$  of the standard-normal distribution (similar as in logistic regression)



# Illustration: Separate MPT Structure for each Person

## Example: 2HTM for two persons

- Probit scores for memory parameter  $d$  are:  $-.10$  and  $1.20$
- What is the predicted probability of correct OLD responses (hits)?
- We assume symmetric and identical guessing for everybody ( $g = .50$ )
- **Person 1:**
  - 1] Transform:  $d = \Phi(-.10) = .46$
  - 2] MPT:  $P(\text{hit}) = d + (1 - d)g = .46 + (1 - .46).50 = .73$
- **Person 2:**
  - 1] Transform:  $d = \Phi(1.20) = .88$
  - 2] MPT:  $P(\text{hit}) = d + (1 - d)g = .88 + (1 - .88).50 = .94$

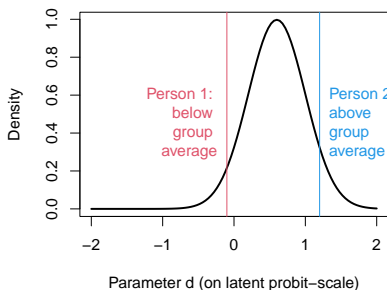


## Group Level: Normal Distribution

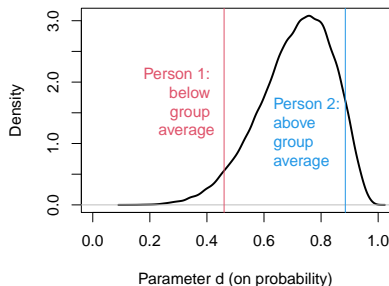
### Assumption: Normal distribution of probit parameters

- Illustration: Normal distribution with mean  $\mu_d = .80$  and standard deviation  $\sigma_d = .3$
- For interpretation, it matters whether parameters are on the probit or the probability scale

Group-Level Distribution (latent probit)



Group-Level Distribution (probability)

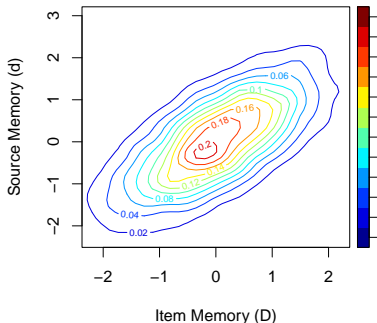
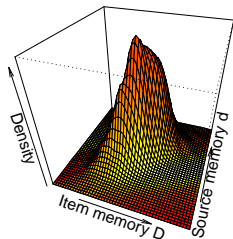




# Comparison of Groups

## Parameter correlations

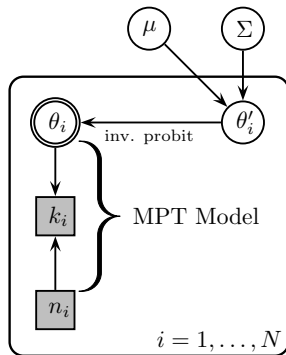
- Item and source memory might be correlated (parameters  $g$  and  $d$ )
- “The more likely I remember the item, the more likely I also remember the source.”
- Solution: Assumption that the vector  $\theta'_i$  with probit-transformed MPT parameters follows a *multivariate* normal distribution
- Caveat: Correlation estimates are often very unprecise and require both large number of responses and large number of participants



# Summary: Hierarchical Models

## Core ideas of hierarchical models

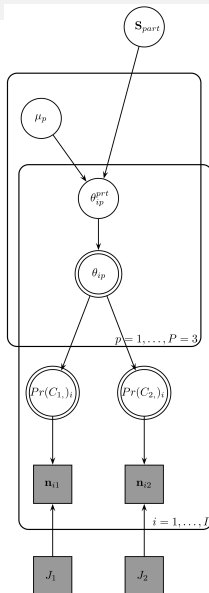
- Assume an MPT model with separate MPT parameters  $\theta_i$  per person
- On the group-level, the parameters have a specific distribution
  - 1] Beta-MPT: Beta distribution
  - 2] Latent-trait MPT: multivariate normal distribution of probit-parameters with mean  $\mu$  and covariance matrix  $\Sigma$
  - 3] Other option (not discussed here): Discrete latent classes (Klauer, 2006)



# Excursion: Graphical Models

## Bayesian graphical models

- In publications, graphical models look more difficult
- Example: Matzke et al. (2015)
- However, most models use exactly the same ingredients



$$S_{part} \sim \text{Scaled-Inverse-Wishart}(\mathbf{W}, df = P + 1, \xi_{part})$$

$$\xi_{part_p} \sim \text{Uniform}(0, 100)$$

$$\mu_p \sim \text{Normal}(0, 1)$$

$$\theta_i^{prt} \sim \text{Multivariate-Normal}\left((\mu_1, \dots, \mu_P), S_{part}^{-1}\right)$$

$$\theta_{ip} = \phi(\theta_{ip}^{prt})$$

$$Pr(C_{11})_i = \theta_{i1} \times \theta_{i2}$$

$$Pr(C_{12})_i = (1 - \theta_{i1}) \times \theta_{i3}^2$$

$$Pr(C_{13})_i = (1 - \theta_{i1}) \times 2 \times \theta_{i3} \times (1 - \theta_{i3})$$

$$Pr(C_{14})_i = \theta_{i1} \times (1 - \theta_{i2}) + (1 - \theta_{i1}) \times (1 - \theta_{i3})^2$$

$$Pr(C_{21})_i = \theta_{i3}$$

$$Pr(C_{22})_i = (1 - \theta_{i3})$$

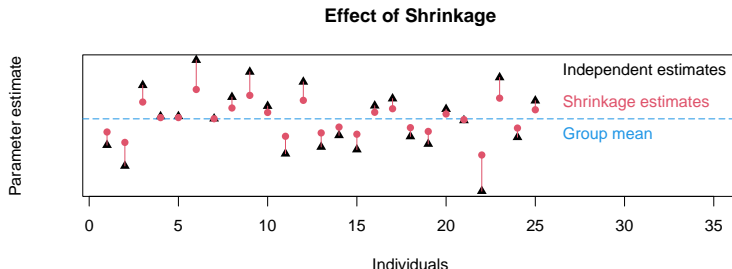
$$\mathbf{n}_{i1} \sim \text{Multinomial}(Pr(C_{1,})_i, J_1)$$

$$\mathbf{n}_{i2} \sim \text{Multinomial}(Pr(C_{2,})_i, J_2)$$

# Some Advantages

## Benefits of hierarchical MPT models

- Avoid aggregation biases
- “Shrinkage” of parameter estimates
  - Parameter estimates for each person are closer together compared to fitting each person separately
  - Hence, extreme estimates are less likely
  - Overall, this ensures that parameter estimates are closer to the true values on average
- The basic idea of hierarchical models can easily applied to any other model
  - 1] Assume that model holds for each person
  - 2] Specify group-level distribution of parameters across persons



## Bayesian estimation with MCMC

## Parameter estimation

- How can we actually fit such models?
- Which are the “best” parameters given the data?
  - Standard MPT models: Maximum likelihood estimation
  - Not an option for hierarchical models (intractable likelihood function due to high-dimensional integrals)

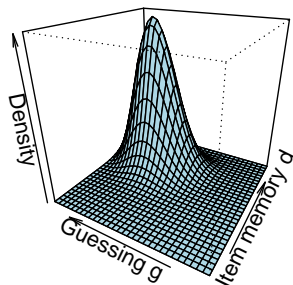
## Solution

- Hierarchical models are often fitted using Bayesian statistics

# Maximum Likelihood

- Logic of parameter estimation with maximum-likelihood
  - 1) Define likelihood function  $p(x | \theta)$
  - 2) Find parameters  $\theta$  that maximize  $f$
- Interpretation: “The estimator  $\hat{\theta}$  has the highest likelihood.”
- Computational solution: search for the “top of the mountain”

## Likelihood



# Bayesian Estimation

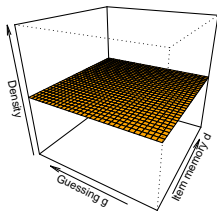
- Logic of Bayesian parameter estimation

- 1 Define likelihood  $p(x | \theta)$  and prior distribution  $p(\theta)$
- 2 Derive the posterior distribution of the parameters via Bayes' theorem:

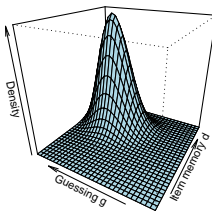
$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

- Interpretation: “What have we learned about the parameters  $\theta$  given the data  $x$ ?”

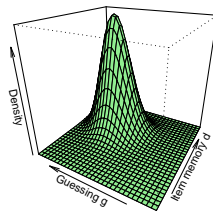
Prior



Likelihood



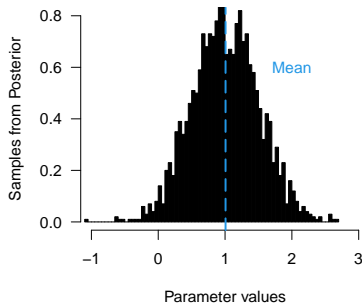
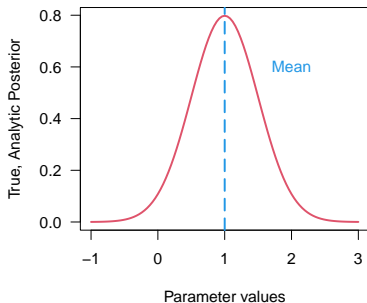
Posterior





# Bayesian Estimation

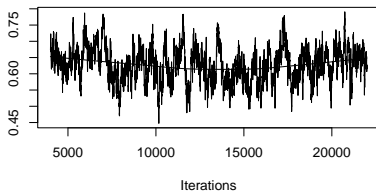
- Problem: We need to work with the posterior function  $p(\theta | x)$ 
  - What is the mean/mode/95% credibility interval of  $\theta$ ?
  - Often, this is analytically not tractable
- Solution: We draw random samples from the posterior distribution
  - Logic: It is easier to draw conclusions from these random samples than deriving solutions for the analytical posterior (which is a function!)
  - Example: Computing the mean of a normal distribution requires to solve:



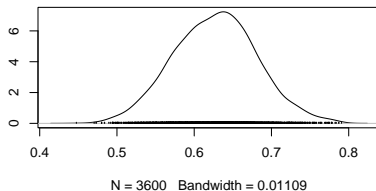
## Markov Chain Monte Carlo (MCMC) Sampling

- 1 Draw random samples of the posterior distribution for *all* parameters (individual and group level)
- 2 Summarize parameter samples (e.g., mean, SD, density, ...)

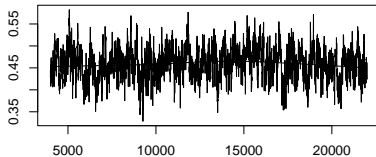
Trace of mean[dn]



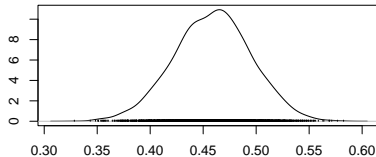
Density of mean[dn]



Trace of mean[g]



Density of mean[g]



## Markov chain Monte Carlo (MCMC)

- General method to draw posterior samples
- In a hierarchical model, there are many (!) parameters
  - Group-level means and covariances, person parameters, ...
  - Intuitively, this method moves around and searches for parameter values with high posterior density
- There are software packages that draw random samples for many models of interest
  - JAGS, WinBUGS, OpenBUGS, Stan, ...

## Summary of Bayesian estimation

- 1) Develop a model ( $\Rightarrow$  psychological theory, multiTree)
- 2) Get posterior (MCMC) samples (JAGS, TreeBUGS)
- 3) Summarize these samples (e.g., mean of group-level parameters  $\mu_D, \mu_g, \dots$ )

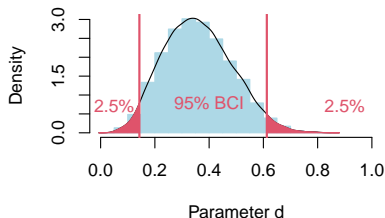
## Advantages of MCMC

# Advantages of MCMC: Uncertainty

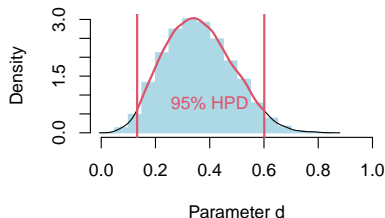
## Advantages of MCMC sampling

- Theoretical:
  - No asymptotic assumptions
  - Maximum likelihood: requires a sufficient number of observations
- Practical: It is easy to quantify uncertainty
  - Bayesian credibility interval (BCI): What are the 2.5%- and 97.5%-quantiles of the parameter values?
  - Highest posterior density interval (HPD or HDI): What are the 95% most plausible parameter values?
  - For probability parameters, these intervals will always be in the interval  $[0, 1]$

**Bayesian Credibility Interval**



**Highest Posterior Density Interval**

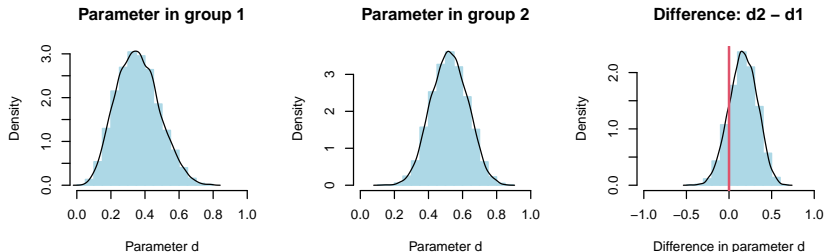


# Advantages of MCMC: Transformed Parameters

- Often, we are interested in parameter/group comparisons
  - Example: Do healthy controls vs. schizophrenics differ in memory?
  - Test: Does the group-mean parameter  $\mu_D$  differ?
- Based on MCMC samples, we can directly estimate functions of the parameters

## MCMC estimation of transformed parameters

- 1 Draw MCMC samples
- 2 Compute transformed parameters for all samples
  - Example:  $\delta^{(t)} = \theta_1^{(t)} - \theta_2^{(t)}$
- 3 Summarize the new values

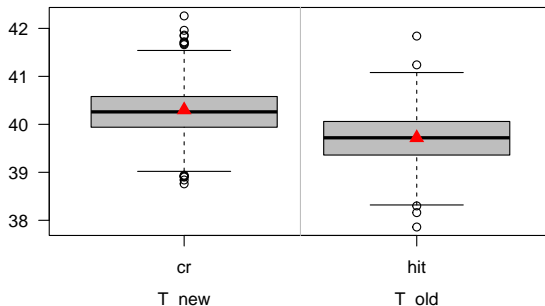


# Advantages of MCMC: Model Fit

## Does the model fit the data?

- Graphical comparison: observed vs. predicted frequencies
  - Use posterior samples of the MPT parameters to sample new data (= posterior predictive)
  - Compare whether these predicted data (boxplot) are in line with the observations (red points)

### Observed (red) and predicted (boxplot) mean frequencies



## How to quantify model fit for MPT models?

- Test statistic similar to Pearson's  $\chi^2$  statistic (Klauer, 2010)
  - T1 statistic: Mean structure of frequencies
  - T2 statistic: Covariance matrix of frequencies
- Posterior predictive  $p$ -value (PPP) measures model fit:
  - 1) Compute T1 for the observed data
  - 2) Compute T1 for the posterior predicted data
  - 3) PPP = probability that T1(predicted) is larger than T1(observed)
- Ideally, PPP should be around .50



## Application

# Linking Covariates to MPT Parameters

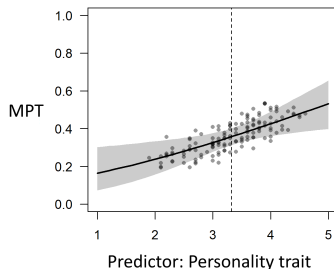
## Interindividual differences

- Personality as a predictor for MPT parameters
- Statistical approach in latent-trait MPT: Similar to logistic regression

$$\theta_i = \Phi(\mu + \boxed{\beta \cdot x_i} + \delta_i)$$

## Cognitive Psychometrics (Riefer et al., 2002)

- Using cognitive (MPT) models to learn about interindividual differences



# Application: Environmental Psychology

## Example: Linking personality to MPT parameters

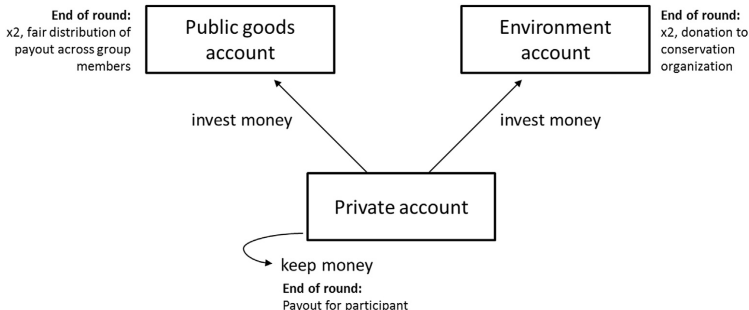
- “Which is the greater good? A social dilemma paradigm disentangling environmentalism and cooperation”
  - Klein, Hilbig, & Heck (2017). *Journal of Environmental Psychology*
- Research question: How can we distinguish between 3 types of behavior?
  - Pro-environmental behavior
  - Pro-social behavior
  - Selfish behavior



# Application: The Greater Good Game

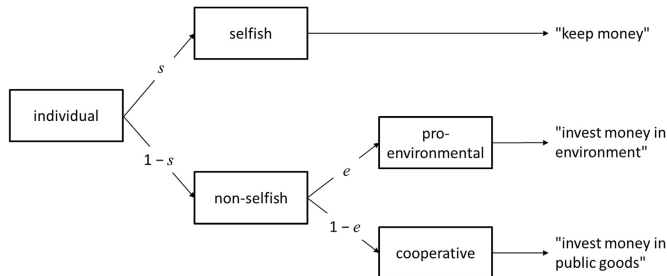
## ■ Greater Good Game

- Participants decide whether to keep the money for themselves or contribute it to either a public goods or an environment account.
- Important: Participants are forced to decide between the group and the environment!
- The game is a variant of a nested public goods game



## MPT model for the Greater Good Game

- $s$  = probability of selfish behavior
- $e$  = probability of pro-environmental behavior



## Results

- Honesty Humility (= sincerity, fairness) is associated with less selfish behavior
- Selfish behavior decreases from 33.4% to 13.9% for participants  $-1/+1$  SD on Honesty Humility

## Hierarchical MPT Models

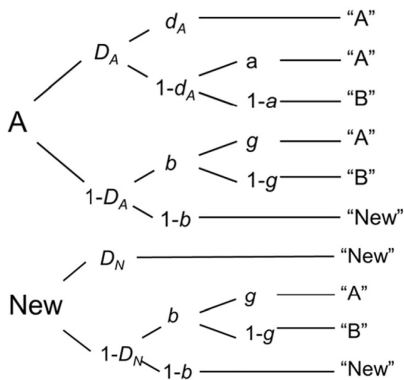
- Individual level
  - Assume a separate MPT model for each person
- Group level
  - Beta-MPT: Beta distribution of person parameters
  - Latent-trait MPT: Normal distribution of probit-transformed parameters
- Bayesian model fitting: Drawing posterior samples

## Appendix & References

# Appendix A: Source-Monitoring Model

## Source-Monitoring

- 1 Study phase: List of words from Source A and B.
- 2 Test phase: Is the presented item from Source A/B/New?



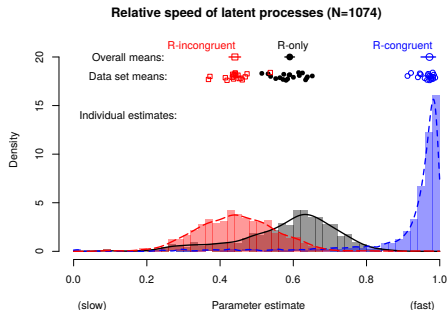


## Appendix B: Meta-Analysis of Raw Data

- Linking process and measurement models of recognition-based decisions (Heck & Erdfelder, 2017, PsychReview)
- Reanalysis of about 400,000 decisions
  - 3-level hierarchical latent-trait MPT:

$$\theta_{sij} = \Phi(\mu_s + \boxed{\xi_{sj}} + \delta_{si})$$

- Overall mean of MPT parameters ( $\mu_s$ )
- Participants nested in studies (random effect:  $\xi_{sj}$ )
- Responses nested in participants (random effect:  $\delta_{si}$ )



## Open questions:

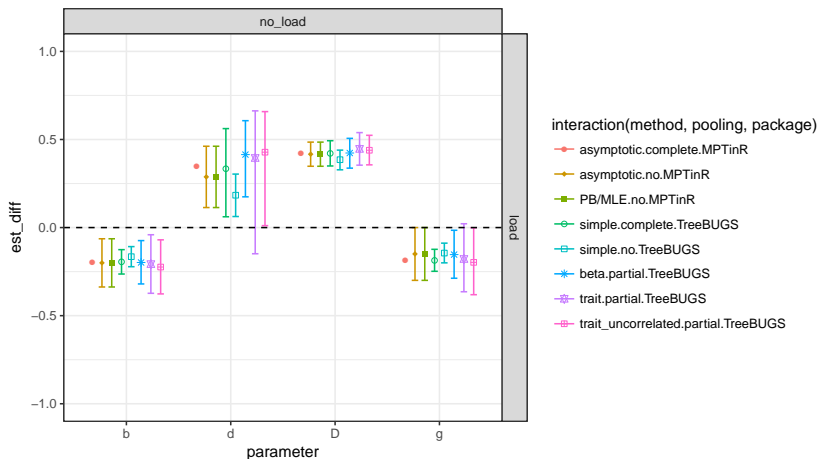
- How much do results actually differ between different MPT versions?
- Which MPT version should be used in practice?

## Large-scale reanalysis project

- Network of MPT researchers (organized by Beatrice Kuhlmann & Julia Groß)
- Reanalysis of existing data sets to compare:
  - Fixed-effects vs. hierarchical
  - Maximum-likelihood vs. Bayes
  - Different hierarchical level-2 structures (beta, multiv. normal, independent univ. normal)
- Software: “A multiverse pipeline for MPT models”
  - Maximum likelihood: `MPTinR` (Henrik Singmann)
  - Bayes: `TreeBUGS`
  - Available at: <https://github.com/mpt-network/MPTmultiverse>

## Appendix C: Reanalysis with Different Models

- Source-monitoring model (data by Bayen & Kuhlmann, 2011)
- Plot: Difference in parameters across two groups



- TreeBUGS and a simple introduction to hierarchical MPT models
  - Heck, D. W., Arnold, N. R., & Arnold, D. (in press). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. Behavior Research Methods. <https://doi.org/10.3758/s13428-017-0869-7>
- The latent-trait model (very technical)
  - Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. Psychometrika, 75, 70–98. <https://doi.org/10.1007/s11336-009-9141-0>
- The latent-trait model with crossed-random effects and a JAGS implementation
  - Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. Psychometrika, 80, 205–235. <https://doi.org/10.1007/s11336-013-9374-9>

- Alternative hierarchical group structure of parameters
  - Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, 54, 167–183. <https://doi.org/10.1016/j.jmp.2009.06.007>
- Benefits of hierarchical cognitive models
  - Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55(1), 1–7. <https://doi.org/10.1016/j.jmp.2010.08.013>
- Cognitive psychometrics
  - Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, 14(2), 184–201. <https://doi.org/10.1037/1040-3590.14.2.184>